Cukurova University
Journal of
Natural & Applied Sciences

Original Paper

# Developing an Automation System for Conflictual Returns Using Machine Learning

**E. Melis Kılıç[1], M. Nezir Alp[2], A. Akyüz Tunç[3], Fatih Abut[4], M. Fatih Akay[5]**

[1,2,3]Trendyol, Data Science Department; Istanbul, Turkey
[4,5]Çukurova University, Faculty of Engineering, Computer Engineering, Adana, Turkey

**ORCID IDs of the authors:** E.M.K. 0000-0002-1625-1221; M.N.A. 0000-0002-0986-9875; A.A.T. 0000-0003-3456-2936; F.A. 0000-0001-5876-4116; M.F.A. 0000-0003-0780-0679.

**Cite this article as:** Kılıç, E.M., Alp, M.N., Tunç, A.A., Abut, F., Akay, M.F. (2022). Developing an Automation System for Conflictual Returns Using Machine Learning. Cukurova University Journal of Natural & Applied Sciences 1 (1): 1-5.

## Abstract

Conflictual returns generate a considerable amount of operational cost in the e-commerce world. Conflictual returns occur in a marketplace when a customer returns a product for certain reasons (broken, missing products, etc.), the seller does not accept the return, and the case becomes unresolved. In the case of conflictual returns, Trendyol needs to resolve the issue as the mediator platform. The decision is made by operators inspecting the customer, the seller, and the case. This process consumes lots of time and human resources. This study aims to automate the resolution of conflictual returns by developing machine learning models based on Logistic Regression (LogReg), CatBoost, and LightGBM. The desired outcome of the study is to make the same decisions on the conflictual returns as the operators as much as possible. The success of the classification models has been evaluated by using the precision, recall, and the area under the curve (AUC)-score metrics. The results show that the proposed LightGBM-based model exhibits the best performance in distinguishing the conflictual returns. The automation of this process will be of great benefit in terms of operational efficiency.

**Keywords:** Data-driven decision making, machine learning, feature selection, e-commerce, classification.

## 1. Introduction

Conflictual returns generate a considerable amount of operational cost in the e-commerce world. Conflictual returns occur in a marketplace when a customer returns a product for certain reasons (broken, missing products, etc.), the seller does not accept the return, and the case becomes unresolved [1-3]. In this case, Trendyol needs to resolve the issue as the mediator platform. The decision is made by operators carefully inspecting the customer, the seller, and the case. However, the detailed inspection of the issue requires time, human resources, and operational costs. Thus, there is a need to automate the decision process of conflictual returns.

A study that specifically tried to come up with a solution to the problem of conflictual returns could not be found in the related literature. However, there are some studies that helped to shape the idea of how to approach and solve the problem. Zhu et al. [4] considered the similarities of products and customers when predicting the return probability of the purchase in e-commerce. Heilig et al. [5] used scalable cloud computing techniques to predict the return probability of the products. Minastirenau et al. [6] used LightGBM model to detect click fraud in online advertising.

Pallathadka et al. [7] discussed the applications of artificial intelligence and machine learning in the e-commerce industry. Moorthi et al. [8] addressed the impact of data analytics techniques in e-commerce and the need for new models and algorithms to collect, store, process, analyze, and evaluate the data in the e-commerce field. A study for automating the decision-making process of

Address for Correspondence:
Fatih Abut, e-mail: fabut@cu.edu.tr

Received: December 5, 2021
Accepted: Jan 7, 2022

conflictual returns is still lacking, and the development of such an automation system can close a significant research gap in the e-commerce industry.

This study aims to automate the resolution of the conflictual returns by developing models based on three machine learning classifiers: Logistic Regression (LogReg), CatBoost, and LightGBM. In more detail, the study plans to automate the process by gathering relevant information about the case, applying machine learning models, and deciding to accept or reject the conflictual return. The contributions of the study can be summarized as follows:

- We propose new models to automate the resolution of conflictual returns by using LogReg, CatBoost, and LightGBM. The desired outcome of the study is to make the same decisions on the conflictual returns as the operators as much as possible.
- We evaluate and compare the performance of all classification models in terms of precision, recall, and AUC-score.
- We rank the applied machine learning classifiers in automating the resolution of conflictual returns.

This paper is structured as follows. Section 2 describes the dataset and methodology. Section 3 presents the results and discussion. Section 4 concludes the paper along with possible future work.

## 2. Dataset and Methodology

Information from all the parties involved is essential to come up with a decision for conflictual returns. Past accepted return rates of the seller, the product, and the customer are essential prediction indicators. The quantity and the price of the product sold are other crucial factors in the decision. The return reason of the customer and the rejection reason of the seller also make a difference in the outcome. These factors are brought together to form the ground-truth dataset. The target metric is whether the conflictual return of the customer should be accepted or not. The target is 0 and 1 for the rejected and accepted returns, respectively. The created dataset consists of 2240 conflictual orders made in February 2021.

To decide in favor of the customer or the seller, various classification models have been developed based on LogReg, CatBoost, and LightGBM. The LogReg gives each predictor a coefficient that measures its independent contribution to variation in the dependent variable [9]. CatBoost and LighGBM are two "Gradient Boosted Decision Tree" implementations. Particularly, CatBoost is a depth-wise gradient boosting library that brings two innovations: Ordered Target Statistics and Ordered Boosting [10]. Similarly, LighGBM contains two novel techniques: Gradient-based One-side Sampling and Exclusive Feature Bundling to handle large number of data instances and large number of features, respectively [11]. Hyperparameter tuning and diverse feature selection and dimensionality reduction techniques like Sequential and Recursive Feature Selection and Principal Component Analysis (PCA) have been applied to achieve the best outcome. Table 1 shows an overview of classification models along with the predictor variables.

The performance of the classification models has been evaluated by using the precision, recall, and AUC-score metrics, the equations of which are given in Eqs. (1), (2), and (3), respectively.

$$precision = \frac{tp}{tp+fp} \tag{1}$$

$$recall = \frac{tp}{tp+fn} \tag{2}$$

$$AUC = \frac{1}{mn}\sum_{m}^{i=1}\sum_{n}^{j=1} 1,\ p_i > p_j \tag{3}$$

Precision is used to define the number of relevant items selected (i.e., the ratio of true positives ($tp$) to all predicted positives ($tp + fp$)). Recall is used to define the number of relevant items retrieved by the supervised model (the ratio of true positives ($tp$) to all actual positives $tp + fn$). The AUC is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the receiver operating characteristic (ROC) curve. In the equation indicating AUC, $m$ represents the data points and $i$ refers to the execution number on $m$ data points that denotes true label. On the other hand, $j$ refers to the execution time of $n$

data points. The equation of AUC in each iteration produces 1 if $p_i > p_j$. $p$ denotes the probability value assigned by the classifier to the related data point [12]. The higher the AUC-score, the better the performance of the model at distinguishing the conflictual returns between "should be accepted" and "should be rejected".

**Table 1.** Overview of classification models along with the predictor variables

| Models | Classifier | Feature Selection / Tuning | Predictor Variables |
|---|---|---|---|
| Model 1 | LogReg | - | Rejected ratio of the product, return ratio of the seller, customer's goodwill score, the quantity and the price of the product sold, the return reason of the customer, the rejection reason of the seller, information about whether the seller has at least 10 orders or not |
| Model 2 | LogReg | Correlation | |
| Model 3 | LogReg | Chi2 - Top n | |
| Model 4 | LogReg | Lasso with Hyperparameter Tuning | |
| Model 5 | LogReg | Sequential Feature Selector | |
| Model 6 | LogReg | Recursive Feature Elimination | |
| Model 7 | LogReg | PCA | |
| Model 8 | CatBoost | Hyperparameter Tuning | |
| Model 9 | LightGBM | Hyperparameter Tuning | |

## 3. Results and Discussion

The results of models in terms of recall, precision, and AUC-score are shown in Table 2. The results are discussed based on model's AUC-scores.

- First, binary classification with LogReg was tried, and the AUC-score was 0.7783.
- Then, the correlation among the features is checked, and feature selection was performed by eliminating the features with more than 0.8 correlation with another feature and a lower correlation with the target than the other feature that is highly correlated with. The results didn't make an uplift compared to the initial trial.
- Then, the chi-square test was applied to the features, and all the features' p-values were smaller than 0.05, which means they are independent features. Feature selection was performed by eliminating the features that have more than 0 p-value. The performance was dropped in this trial; however, it was more successful than the first two trials, deciding the top 50% probability orders in the simulation.
- In this trial, the Lasso algorithm in LogReg was applied by making penalty = L1.
- Different values of C were tried and the value of C leading to the highest AUC-score was selected. AUC-score jumped to 0.9036.
- Sequential feature selection was applied, and the "number of features" parameter was tuned to get the maximum AUC-score. The AUC-score increased to 0.9096.
- Recursive feature selection was applied, and the "number of features" parameter was tuned to get the maximum AUC-score. Recursive feature selection didn't increase the performance.
- PCA was applied to the features, and this didn't bring an extra uplift.
- The CatBoost-based model was applied. Tree depth and the number of trees hyperparameters were tuned to maximize the AUC-score, and the AUC-score increased to 0.9374.
- The LightGBM-based model was tried, and the hyperparameters were also tuned to maximize the AUC-score, which slightly increased to 0.9378.

As a result of the different model type trials, feature selection methods, and hyperparameter tunings, the LightGBM-based model was the most successful version with the maximum AUC-score.

Table 3 shows the simulation results of different model iterations on past conflictual return data. The orders were sorted by the probability of being accepted in descending order, divided into ten chunks, and the overall realized acceptance rates were calculated. With the LightGBM-based Model 9 that gave the highest AUC-score, the first seven chunks' acceptance rate is near 100%, and the last chunk's acceptance rate is near 0%. It can be concluded that we can safely automate the decision-making process of at least 80% of the conflicted returns.

**Table 2.** The performance of models in terms of precision, recall, and AUC-score

| Models | Precision | Recall | AUC-Score |
|---|---|---|---|
| Model 1 | 0.9128 | 0.9914 | 0.7783 |
| Model 2 | 0.9119 | 0.9885 | 0.7731 |
| Model 3 | 0.8385 | 0.9956 | 0.5422 |
| Model 4 | 0.9629 | 0.9802 | 0.9036 |
| Model 5 | 0.9665 | 0.9799 | 0.9096 |
| Model 6 | 0.9622 | 0.9807 | 0.9034 |
| Model 7 | 0.9659 | 0.9787 | 0.9079 |
| Model 8 | 0.9767 | 0.9829 | 0.9374 |
| Model 9 | **0.9778** | **0.9969** | **0.9378** |

**Table 3.** Simulation results of different model iterations on past conflictual return data

| Probability Rank | Acceptance Rates | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Model 1 | Model 2 | Model 3 | Model 4 | Model 5 | Model 6 | Model 7 | Model 8 | Model 9 |
| top 10% | 0.9996 | 0.9991 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 10-20 | 0.9987 | 0.9987 | 0.9996 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 | 1.0000 |
| 20-30 | 0.9987 | 0.9987 | 1.0000 | 1.0000 | 0.9996 | 0.9996 | 0.9996 | 1.0000 | 1.0000 |
| 30-40 | 0.9929 | 0.9937 | 0.9991 | 0.9991 | 0.9987 | 0.9982 | 0.9973 | 0.9996 | 0.9996 |
| 40-50 | 0.9853 | 0.9870 | 0.9915 | 0.9969 | 0.9973 | 0.9960 | 0.9978 | 1.0000 | 0.9996 |
| 50-60 | 0.9777 | 0.9665 | 0.9763 | 0.9915 | 0.9879 | 0.9879 | 0.9906 | 0.9982 | 0.9969 |
| 60-70 | 0.9580 | 0.9455 | 0.9553 | 0.9741 | 0.9754 | 0.9781 | 0.9750 | 0.9826 | 0.9951 |
| 70-80 | 0.8450 | 0.8330 | 0.8312 | 0.8749 | 0.8991 | 0.8678 | 0.8888 | 0.9227 | 0.9272 |
| 80-90 | 0.3939 | 0.4256 | 0.3912 | 0.3528 | 0.3765 | 0.3417 | 0.3770 | 0.3064 | 0.3394 |
| 90-100 | 0.0607 | 0.0768 | 0.1166 | 0.0192 | 0.0205 | 0.0174 | 0.0214 | 0.0040 | 0.0054 |

## 4. Conclusion and Future Work

This study aimed to save human resources by automating the decision-making process of conflictual returns, which is currently made by the human mind. To come up with a decision for conflictual returns, various classification models based on LogReg, CatBoost, and LightGBM were developed by using data related to the conditions of the customer, the seller, and the order. The results show that the LightGBM-based model with an AUC-score of 0.9378 provides satisfying performance in distinguishing the conflictual returns. The automation of this process will be of great benefit in terms of operational cost and efficiency. The simulation results show that more than 80% of the conflictual returns could automatically be resolved by our proposed LightGBM-based model.

For the next steps, more potential predictors of conflictual return resolutions, such as the type of purchase (e.g., food, clothing, and electronic device), can be evaluated to investigate whether the accuracy of classification models can be improved. In addition,

the utilized dataset can be enriched with additional conflictual return samples. Other classifiers, such as Long Short-Term Memory, Support Vector Machine, and Multiplayer Perceptron, can be evaluated to increase the classification accuracy. Ultimately, we desire to develop a system to conclude all the decisions in real-time automatically.

## References

[1] Lin, D., Lee, C.K.M., Siu, M.K., Lau, H. and Choy, K.L. (2020). "Analysis of customers' return behaviour after online shopping in China using SEM", Industrial Management & Data Systems, vol. 120, no. 5, pp. 883-902.

[2] Jadhav, A. (2021). "Effect of Sales Returns on Retail and E-commerce Industry". Available at Social Science Research Network (SSRN): https://ssrn.com/abstract=3833921 or http://dx.doi.org/-10.2139/ssrn.3833921.

[3] Han, H. (2019). "Review and Prospect on Return Problems of E-Commerce Platform". Open Journal of Business and Management, vol. 7, pp. 837-847.

[4] Zhu, Y., Li, J., He, J., Quanz, B. L. and Deshpande, A. A. (2018). "A Local Algorithm for Product Return Prediction in E-Commerce". In: Proc. of the International Joint Conference on Artificial Intelligence, pp. 3718-3724.

[5] Heilig, L., Hofer, J., Lessmann, S. and Voß, S. (2016) "Data-Driven Product Returns Prediction: A Cloud-Based Ensemble Selection Approach". In: Proc. of the 24th European Conference on Information Systems, pp. 1-10.

[6] Minastireanu, E. A. and Mesnita, G. (2019). "Light GBM Machine Learning Algorithm to Online Click Fraud Detection". Journal of Information Assurance & Cyber security, Article ID 263928.

[7] Pallathadka, H., Ramirez-Asis, E. H., Loli-Poma T. P. et al. (2021) "Applications of artificial intelligence in business management, e-commerce and finance", Materials Today: Proceedings, https://doi.org/10.1016/j.matpr.2021.06.419.

[8] Moorthi, K., Dhiman, G., Arulprakash P. et al. (2021). "A survey on impact of data analytics techniques in E-commerce", Materials Today: Proceedings, https://doi.org/10.1016/j.matpr.2020.10.867.

[9] Boateng, E. and Abaye, D. (2019). "A Review of the Logistic Regression Model with Emphasis on Medical Research". Journal of Data Analysis and Information Processing, vol. 7, pp. 190-207.

[10] Hancock, J. T., Khoshgoftaar, T. M. (2020). "CatBoost for big data: an interdisciplinary review". Journal of Big Data, vol. 7, no. 1, p. 94.

[11] Ke, G., Meng, Q., Finley, T. et al. (2017). "LightGBM: a highly efficient gradient boosting decision tree". In: Proc. of the 31st International Conference on Neural Information Processing Systems. New York, USA, pp. 3149–3157.

[12] Öztürk, M. M. (2019). "The impact of parameter optimization of ensemble learning on defect prediction". Computer Science Journal of Moldova, vol. 27, no. 1, pp. 85-128.